All in the Family? A Co-Authorship Analysis of JCDL Conferences (1994-2003)

Xiaoming Liu[†] Johan Bollen[‡]
[†] Research Library
Los Alamos National Laboratory
Los Alamos, NM, 87545 USA
{liu_x,herbertv}@lanl.gov

Michael L. Nelson [‡] Herbert Van de Sompel[†]
[‡] Computer Science Department
Old Dominion University
Norfolk, VA, 23529 USA
{jbollen,mln}@cs.odu.edu

ABSTRACT

The field of digital libraries (DLs) coalesced in 1994: the first digital library conferences were held that year, the awareness of the World Wide Web was accelerating, and the National Science Foundation awarded $24 Million (U.S.) for the Digital Library Initiative (DLI). In this paper we examine the state of the DL domain after a decade of activity by applying social network analysis to the co-authorship network of the past ACM, IEEE, and joint ACM/IEEE digital library conferences. We base our analysis on a common binary undirected network model to represent the co-authorship network, and from it we extract several established network measures. We also introduce a weighted directional network model to represent the co-authorship network, for which we define *AuthorRank* as an indicator of the impact of an individual author in network. We show that the weighted directional model allows for the deployment of an author navigator application for convenient visual examination of the co-authorship network. We also investigate the amount and nature of international participation in JCDL.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

1. INTRODUCTION

In 1994, the National Science Foundation awarded $24 Million (U.S.) to six institutions, thereby officially kicking off the federally-sponsored DL research program. Also in 1994, the first of what was later to become the IEEE Advances in Digital Libraries (ADL) conference and the ACM Digital Libraries (DL) conference were held in New Jersey and Texas, respectively. In 2001, the two conference series were merged and the first ACM/IEEE Joint Conference on Digital Libraries (JCDL) was held in Virginia. These conferences have induced a pattern of collaborations which has shaped the domain of DLs over the past decade. To study the structure of these collaborations, we used social network analysis to investigate authorship trends in the composite corpus of the long papers, short papers and posters presented at all those DL conferences.

Social network analysis has attracted considerable interest in recent years and plays an important role in many disciplines [4]. In information science, social network analysis has applications in citation networks, co-citation networks, collaboration networks, and recently also in web graph analysis [19, 24].

A co-authorship network is a type of collaboration network. A popular culture example is the Oracle Of Bacon project [23], which determines the distance between any actor and Kevin Bacon by examining movie co-starring relationships. An early example of analyzing a scholarly co-authorship network is the Erdös Number Project, in which the smallest number of co-authorship links between any individual mathematician and the Hungarian mathematician Erdös were calculated [8]. Newman studied and compared the co-authorship graph of arXiv, Medline, SPIRES, and NCSTRL [18]. Co-authorship analysis has also been applied to various ACM conferences: Information Retrieval (SIGIR)[22], Management of Data (SIGMOD) [17] and Hypertext [10], as well as mathematics and neuroscience [13] and information systems [11]. Although the above projects have successfully observed interesting patterns in specific communities, many of them focus on special features of the observed network. Furthermore, most network models used in those analysis are undirected and have a binary (0 or 1) weight for the edges between authors.

In this paper we present a study of the co-authorship network of DL conferences (ACM DL, IEEE ADL, and JCDL) held between 1994 and 2003. A binary undirected co-authorship network is built using XML encoded data available from DBLP [15]. Social network metrics, including small world analysis, component analysis, and centrality analysis, are applied to this network.

Furthermore, we extend this binary network by considering co-authorship frequency. A long-time collaboration between two researchers - resulting in many co-authored papers - might be considered more important than an occasional co-authorship in a large cross-institutional project - resulting in many co-authors in a single paper. Thus we propose a weighted directed network model to represent the co-authorship network, and AuthorRank, an alternative centrality metric. Several of the presented centrality metrics are then cross-validated against the dataset of ADL/DL/JCDL program committee members. Also AuthorRank rankings obtained from the directed weighted network model are correlated with metrics obtained from the undirected binary network model. We show that AuthorRank can also be used by an author navigator for convenient lookup of the co-authorship network. We also investigate the amount and nature of international participation in JCDL. The presented material sketches an interesting picture of the JCDL community, and the approach used can also be easily extended to other collaboration networks.

The remainder of this paper is organized as follows: Section 2 discusses related work; in Section 3 we describe the network models and metrics used in our analysis; Section 4 presents the results of our analysis of the JCDL co-authorship graph using those models and metrics, and Section 5 presents our conclusions.

2. RELATED WORK

Social networks have been long studied and have gained increasing importance in recent years [24]. Social network analysis studies relationships among social entities - *actors*, and the patterns and implications of these *relationships*.

Various metrics are applied in social networks analysis to study either global properties or actor properties. For global properties, the whole network is analyzed, and attempts are made to find all relations by components, cliques, small worlds, etc. For actor properties, the network context of a single actor is analyzed, and attempts are made to measure the differences between important and non-important actors. The importance of an actor can be measured by centrality - i.e. position of the actor in the network - or prestige - endorsements received by the actor from other actors. Prestigious actors are not only endorsed by many actors, but the endorsing actors must also be prestigious. This recursive nature of prestige is mathematically addressed by eigenvector analysis. Basic eigenvector analysis is sometimes wrongly applied to asymmetric networks in which some actors do not receive any endorsement. Enhancements such as alpha-centrality have been developed to address this problem [7]. Eigenvector analysis is also used to measure the prestige of web pages; well-known algorithms include Google's PageRank [21] and Kleinberg's HITS [14]. However, in both algorithms all edges have binary weight. Bharat and Henzinger [6] developed a weighted edge scheme to improve the HITS algorithm, and it is also possible to modify the PageRank formulation to take edge weight into account [9]. In collaboration networks, Li and Chen used numbers of collaborations to define a weighted network model [16]. They concluded that the distribution of connection strengths decays in a power-law form.

3. FRAMEWORK AND MODEL

Graph theory has been used in social network analysis since the late 1940s [24]. Thereby, the whole network is treated as a *graph*, consisting of *nodes* joined by *edges* or *links*. In this paper, the terms *node*, *actor*, and *author* are interchangeable. Simiarly, the terms *edge*, *relationship*, and *co-authorship* are also used interchangably.

3.1 Binary Undirected Co-Authorship Network

Perhaps the simplest co-authorship network model is based on an undirected graph G in with each edge represents a co-authorship relationship. Consider two articles:

article 1, Author v_1*, Author* v_2*, Author* v_3
article 2, Author v_1*, Author* v_2

If any two authors co-authored an article, an edge with unit weight is created (Figure 1). The graph is denoted as a undirected unit-weighted graph $G = (V, E)$, where the set of n authors is denoted $V = \{v_1, ...v_n\}$ and E represents the edges between authors. As will be shown in the next sections, various graph metrics can be extracted from this kind of network.

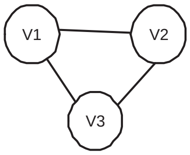

Figure 1: Binary undirected co-authorship network

3.2 Metrics for a Binary Undirected Co-Authorship Network

A number of social network metrics are available for measuring the characteristics of a binary undirected collaboration network, including components analysis, small world analysis, and centrality analysis. These metrics measure various network properties and some may only be applied under certain conditions. The metrics used in this paper are listed in Table 1 and discussed below.

3.2.1 Component Size Analysis

A component of a graph is a subset with the characteristic that there is a path between any node and any other node of this subset. A co-authorship network usually consists of many disconnected components (e.g. disconnected research groups or individuals), and component analysis can be used to learn about the structure of the network. Some network analysis methods are only widely used in connected networks. Therefore, in networks with disconnected components, those methods are typically only applied to the largest connected component, as shown in Table 1.

3.2.2 Small World Analysis

Small world analysis is typically only applied in a fully connected network, and is used to measure global properties of the network. The characteristic path length of a graph G is defined as the average shortest path length between each pair of vertices [25] . The clustering coefficient measures how well the direct neighbors of a vertex are connected among themselves. A graph is a small world graph if it is characterized by the following two conditions: (1) it has a much higher clustering coefficient than a similarly sized random graph, and (2) it has only a slightly larger characteristic path length than a similarly sized random graph. Studies have shown that many collaboration networks are small world graphs [5].

Table 1: Co-authorship network metrics

Metric	Type		Property		Scope		Importance	
	Binary	Weighted	Actor	Global	Whole Network	Largest Component	Centrality	Prestige
Component	×			×	×			
Small World	×			×		×		
Cluster		×		×		×		
Closeness	×		×			×	×	
Betweeness	×		×		×		×	
Degree	×		×		×		×	
PageRank	×		×		×			×
AuthorRank		×	×		×			×

3.2.3 Degree, Closeness, Betweenness Centrality

The authoritative text by Wasserman [24], defines three centrality measures commonly used in social network analysis: degree centrality, closeness centrality, and betweenness centrality. Degree centrality of a node is defined as the number of edges adjacent to this node. The closeness centrality of a node is equal to the total distance of this node from all other nodes. Since a shorter distance means better connectivity, standardized closeness, which is the inverse measure of distance, is used to measure centrality. Closeness centrality is best used in connected networks. Betweenness centrality is the number of shortest paths that pass through a given node. Betweenness centrality can be used in disconnected networks, however it may generate a large number of nodes with zero centrality, since many nodes may not act as a bridge in the network, as shown in Table 1.

3.2.4 PageRank

PageRank is the ranking mechanism at the heart of Google. It can either be explained by a link-based analysis or by a random walk model [20, 21]. PageRank forms a probability distribution over web pages. PageRank can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. PageRank is originally designed to measure the hyperlink structure of the web, which is a directed graph in nature. In order to use PageRank in the context of a co-authorship network, we build two directional edges e_{ij} and e_{ji} for each edge e between author v_i and v_j. This fits well with the two models underlying PageRank: in the link analysis model, the directional edges can be understood as the mutual endorsement of authors; in the random walk model, the directional edges can be understood as the movement of a surfer in either direction on edge e. PageRank can legitimately be applied to the proposed network model since it is a special case of the network model underlying the PageRank metric.

3.3 Weighted Co-Authorship Network

We believe that the binary undirected network without link weights does not fully represent frequency and importance of interaction in a co-authorship network. There are many cases in which the binary undirected network does not correspond with a common sense notion of magnitude. For example, if two authors co-publish many papers, should the link between them be considered more important than the link between occasional co-authors? Also, if one article has two authors and another article has a hundred authors, should the authors in the first article be considered more connected than those of the second article? Although these assumptions are arguable, we consider a link between two authors to be more important than another if: (1) they have co-authored more papers, or (2) their papers have fewer co-authors.

Therefore, we formally define several metrics to measure the importance of links. Let the set of n authors be denoted as $V = \{v_1, ...v_n\}$. Let the set of m articles be denoted as $A = \{a_1, ..., a_k, ...a_m\}$, and $f(a_k)$ is the number of authors of article a_k. We define:

Weight of Link in Single Article: If author v_i and v_j are co-authors in article a_k,

$$g(i, j, k) = 1/(f(a_k) - 1) \tag{1}$$

This gives more weight to co-author relationships in articles with fewer total co-authors than articles with large numbers of co-authors.

Accumulated Weight of Link:

$$c_{ij} = \sum_{k=1}^{m} g(i, j, k) \tag{2}$$

This gives more weight to authors who co-publish more papers together.

Normalized Directed Weight:

$$w_{ij} = c_{ij} / \sum_{k=1}^{n} c_{ik} \tag{3}$$

This ensures that the weights of all of an author's relationships sum to one.

We represent the co-authorship network as a directed weighted graph. A graph G is denoted $G = (V, E, W)$, where V is the set of nodes (authors), E is the set of edges (co-author relationships between authors), and W is the set of weights w_{ij} associated with each edge connecting a pair of authors (v_i, v_j).

Similar to PageRank, the weighted co-authorship network model also has an intuitive basis in random walks on graphs (Figure 2). The normalized weight corresponds to the probability distribution of a random walk on the co-authorship graph. A random walker may choose to start navigating the network from any author. In Figure 2, if the walk starts from author v_1, the walker may travel to v_2 or v_3 with probability

0.75 and 0.25 respectively. If the walker starts from author v_3, however, the walker has same probability of visiting v_1 or v_2. The weighed co-authorship also has an intuitive meaning as the endorsement of an author. For example, from Figure 2, we can understand that v_1 and v_2 have a higher mutual endorsement since they co-authored more papers.

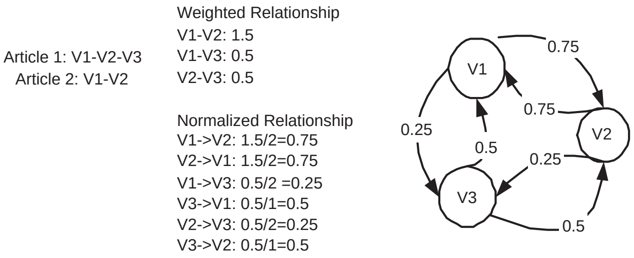

Figure 2: Weighted co-authorship network

3.4 Metrics for a Weighted Co-Authorship Network

3.4.1 AuthorRank

Similar to PageRank, we assume Authors $v_{t1}...v_{tn}$ pointing to author v_i. We also use the same dampening factor d in PageRank, and the weight w is defined in the previous section. The AuthorRank of an author v_i is given as follows:

$$AR(v_i) = (1-d) + d(AR(v_{t1}) * w_{v_{t1},i} + ... + AR(v_{tn}) * w_{v_{tn},i}) \quad (4)$$

The AuthorRank can be calculated with the iterative algorithm used by PageRank. One may think of AuthorRank as a generalization of PageRank by substituting $w_{v_{tn},i}$ with $1/number\_of\_outlink$ in PageRank.

Looking at the examples underlying Figure 1 and Figure 2 reveals an obvious advantage of AuthorRank (defined for the weighted network) over the centrality measures and PageRank (defined for the binary network): the latter will generate same rank for v_1, v_2, v_3 which is counter-intuitive as we argued before. AuthorRank, however, will generate the same rank for v_1 and v_2, and a lower rank for v_3.

3.5 Interactive Author Navigator

The weighted model also has an advantage for the visualization of a co-authorship graph. The visualization of a well connected author can be rather crowded and difficult to read. With a weighed network model, important links can be emphasized and trivial links can be truncated. As one can imagine, the whole co-authorship graph is too large to fit on a single computer screen. To remedy this, we built an interactive author navigation tool based on the webdot tool of GraphViz [2]. Users can select a preferred author (center of the graph), set a distance from the selected author, and indicate the minimum weight necessary for links to be displayed. Based on those parameters, a subgraph is dynamically constructed and visualized. In this visualization, the weight of a link plays an important role as it allows users to identify important links. The graph is clickable and the user can navigate to other interesting authors.

4. JCDL CO-AUTHORSHIP ANALYSIS

4.1 Generating the Co-authorship Network

We extracted co-authorship data from DBLP for ACM DL (1995-2000), IEEE ADL (1994-2000), and JCDL (2001-2003). This includes all long papers, short papers, posters, demonstrations, and organizers of workshops.[1] The dataset contained 1567 authors, 759 publications, and 3401 co-authorship relationship pairs. Some statistics are readily available from this data set. For example, the number of articles, authors, international (non-US) authors, and new authors per year is shown in Figure 4. It can be seen that number of articles and the number of authors are highly correlated, and that a major boost occurred following the merger of the ACM/IEEE DL series into a single JCDL conference. Figure 3 shows the number of publications per author. The values range between 1 and 22, with 4 authors publishing more than 10 papers and 78% of the authors publishing only 1 paper and 95% authors having 3 papers or less. Authors with 8 or more publications are shown in Table 2. Each paper has a mean of 3.02 authors and a median of 3 authors. The distribution of number of authors per paper is shown in Table 3.

We also studied international collaboration. Approximately 72% (1133/1567) of the authors are affiliated with U.S. institutions. We discovered that among 3401 co-authorship relationships, only about 7% are collaborations between authors from different countries. A country collaboration network is created by accumulating cross-country collaborations from the author network. In Figure 5, we represent countries by country code domain [1]; two countries are closer to each other if authors from those countries collaborated closely. The figure can only be considered approximate due to the limitations of the visualization technology used. Figure 5 shows that the JCDL community is centred around .us, with .uk, .nz, and .sg closely surrounding .us; .nz and .de also play significant roles in connecting different countries. There are nine countries (.es, .ie, .at, .hu, .nl, .in, .kr, .il, and .za; with 61 authors) that are not connected with other countries. The distribution of authors from each country is shown in Figure 6.

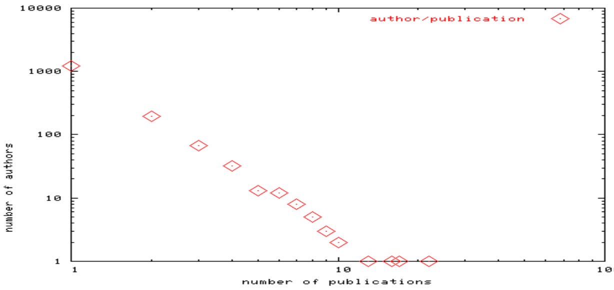

Figure 3: Number of papers per author

[1] Unfortunately, we discovered just prior to submission that DBLP had mislabeled ADL 94 as DL 94 and omitted DL 94 altogether. We are entering the data for DL 94 manually, and will recalculate our results. However, we do not believe that the inclusion of DL 94 will significantly alter our findings.

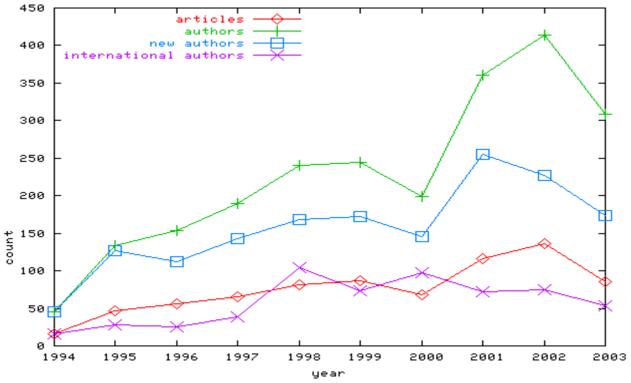

Figure 4: Articles, authors, international authors, and new authors per year

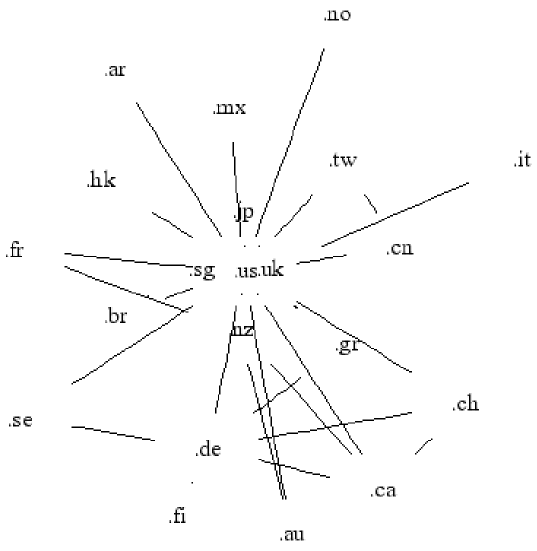

Figure 5: Country network

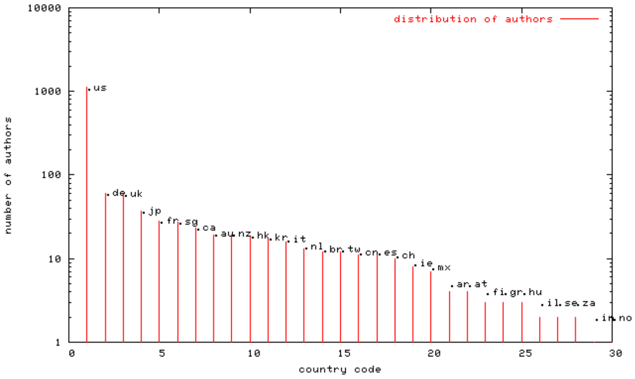

Figure 6: Distribution of authors per country

Table 2: Authors with 8 or more publications

Name	Publications
Hsinchun Chen	22
Edward A. Fox	17
Ian H. Witten	16
Hector Garcia-Molina	13
Alexander G. Hauptmann	10
Gary Marchionini	10
Judith Klavans	9
Carl Lagoze	9
Michael L. Nelson	9
David Bainbridge	8
Richard Furuta	8
Ee-Peng Lim	8
Catherine C. Marshall	8
Terence R. Smith	8

Table 3: Distribution of number of co-authors per paper

Number of authors	Number of papers	Percentage
1	149	19.6%
2	216	28.5%
3	179	23.6%
4	94	12.4%
5	45	5.9%
6	33	4.3%
7	20	2.6%
8	7	0.9%
9	4	0.5%
10	5	0.7%
11	1	0.1%
12	2	0.3%
13	1	0.1%
14	1	0.1%
15	2	0.3%
total		100%

4.2 Component Size Analysis

As is to be expected, the co-authorship network is not a single connected graph. The largest component of the network has 599 authors, the second largest component has 31 nodes and so on. The distribution of component size roughly follows a power law distribution as can be seen in Figure 7. The entire co-authorship network with all components is shown in Figure 8, in which nodes represent authors and links represent collaborations. The largest component is on the left side of the Figure, while the right side shows many small components. Well-connected components are recognizable by their very dense (dark) shape.

Nascimento [17] reports that the largest component in SIGMOD's co-authorship graph has about 60% of all authors. In the four co-authorship networks studied by Newman [18], NCSTRL has the smallest largest component,

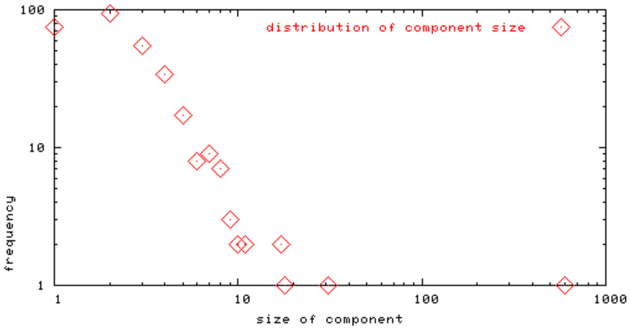

Figure 7: Distribution of component size

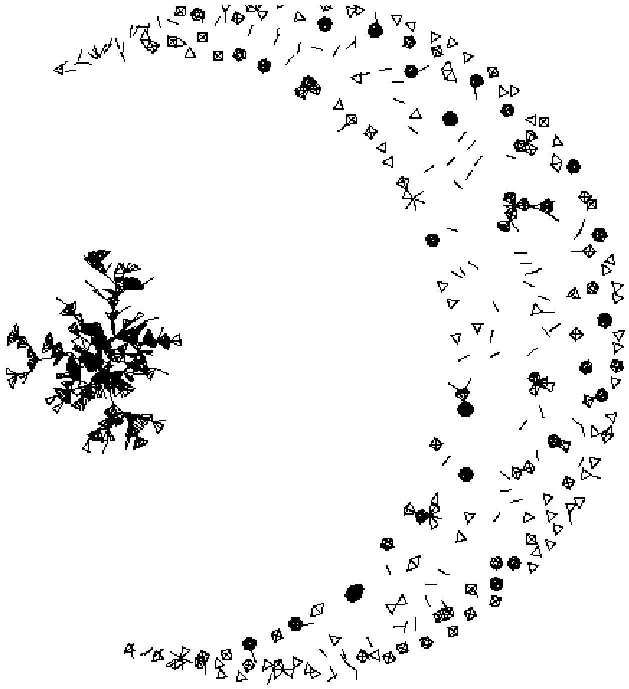

Figure 8: Component size analysis

containing 57.2% of all authors. However, in the JCDL co-authorship network the largest component only includes 38% of all authors. Several possible explanations could account for this low value, including the relative immaturity of the the DL field, the multi-disciplinary nature of the composite JCDL conference series, the fact that many DL projects grow from a "grass-roots", institutionally oriented focus[12], or the rather restricted nature of international collaboration in the DL field.

To better understand the nature of major components and the reason for them not being in the large component, we conducted a manual analysis of other large components. This showed that the most dense shapes include authors from the same institution or working on the same project. We counted 18 components with sizes ranging from 7 to 31. By checking the affiliation of authors, we discovered that 5 components consist mainly of non-US participants, and that the 31-node component represents the medical informatics community. By checking titles and content, we found that 13 components account for short papers or posters only, many of which are about a specific DL application in a particular scenario. Therefore, it is our guess that the short paper and poster programs encourage a wide participation from other disciplines.

4.3 Cluster Analysis

The weighted graph model also improves the clustering because close and frequent collaboration causes higher similarity scores between authors, resulting in them being grouped closer together. By representing each author as a vector of relationships to other authors using the weighted graph model, we conducted a bottom-up, hierarchical clustering algorithm on the largest component of the co-authorship network. The hierarchical clustering algorithm starts with all authors and successively combines them into groups with high inter-authorship similarity. Typically, the earlier mergers happen between groups with a large similarity, and similarity becomes lower and lower for later merges. The result reveals initial clusters do not necessarily reflect institutional boundaries. This may be due to the fact that authors may change institutions, and in some cases strong collaborations exist between institutions. In the next stage institutions are merged into larger clusters due to their joint publications or common research interests. A well-connected author is usually only clustered in this stage, which confirms that well-connected authors play an important role in connecting different clusters.

As a matter of illustration, the clusters to which the authors of this paper belong are shown in Figure 9. As can be seen, small clusters are initially formed in each authors' institution (LANL and Old Dominion Univerity), and later institutions are merged to larger clusters. The frequency of joint publications may explain the different stage of merging. By checking publications in each cluster, we found that LANL, Cornell University and the University of Southampton form a larger cluster because Cornell cooperated with Southampton in the Open Citation project, and LANL worked with Cornell on the Open Archives Initiative. Similarly, Virginia Tech (VT) collaborated with the Federal University in the Web-DL project, with Penn State (PSU) in the CITIDEL project, and with Old Dominion University (ODU) in the NCSTRL project. ODU and PSU have no joint publications, they are clustered together because both collaborated with VT. VT and Federal University probably merged earlier because they have more joint publications.

4.4 Small World Analysis

Since small world analysis can only be done in a connected graph, we used the largest component of the co-authorship network for our calculation. The largest component (599 authors and 1897 links) has a clustering coefficient of 0.89, and characteristic path length of 6.58. With a similarly sized connected random graph, the clustering coefficient is 0.31 and characteristic path length is 3.66. This means that the JCDL co-authorship network is indeed a small world graph. The largest component is shown in Figure 10.

Nascimento [17] reported that the SIGMOD co-authorship graph yields a clustering coefficient of 0.69, and a characteristic path length of 5.65. In all four networks studied by Newman, the largest clustering coefficient generated is 0.726. This shows a rather high clustering coefficient of the JCDL co-authorship network, meaning that co-authors of one author are more likely to publish together. The JCDL co-authorship network also has a rather long characteristic path length, indicating that authors from different groups are not as well connected as, for example, those in the SIGMOD co-authorship network.

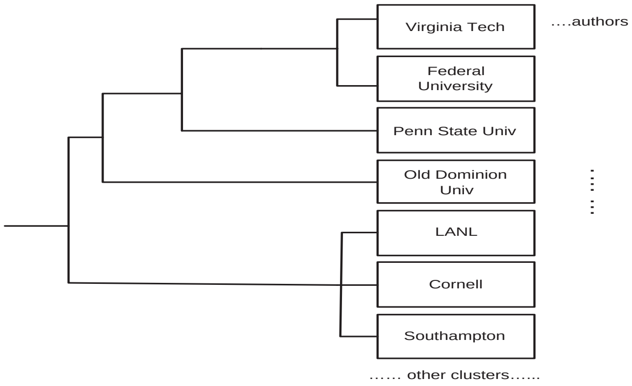

Figure 9: Clustering result

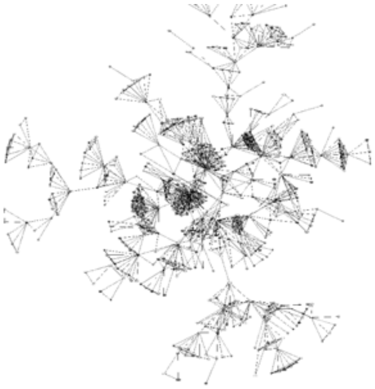

Figure 10: Largest component of the co-authorship network

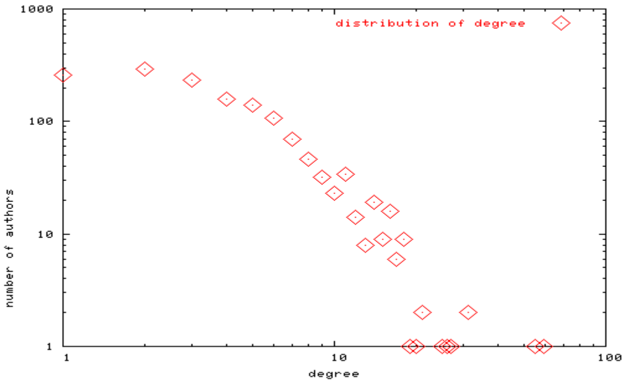

Figure 11: Degree distribution

4.5 Centrality

Using the R package [3], we calculated the degree, closeness, and betweenness centrality for the binary undirected co-authorship network only, as these metrics are not well defined in a weighted network. The highest ranking 20 authors for each metric and their scores are listed in Table 4.

4.5.1 Degree Centrality

The degree centrality distribution is shown in Figure 11. It also follows a rough power-law distribution with few authors having a high degree of connection. This measurement has the disadvantage of giving many authors the same weight. It is also biased to authors with many collaborators in a single publication.

4.5.2 Closeness Centrality

The closeness centrality is only applied to the largest component (599 authors) since closeness is not well defined in a disconnected network. It has a bias towards authors that are directly connected to a well-connected author. For example, a graduate assistant of a prestigous professor may have a fairly high weight.

4.5.3 Betweenness Centrality

The betweenness centrality is applied to the whole network, however only 153 authors have positive values. The remaining 1414 authors do not lie on the paths between other authors. The computation of betweeness centrality is the most resource-intensive of all measures we explored, since it requires enumerating all of the shortest paths between each pair of nodes. It takes more than one day to calculate betweenness on an Intel Celeron machine using R [3].

4.6 PageRank and AuthorRank

We developed a Java program with a MySQL backend to calculate PageRank and AuthorRank. Both calculations can be completed in several seconds. The 20 highest scoring authors for the PageRank and AuthorRank metrics are listed in Table 5.

4.7 Correlation and Validation

Several articles have compared the performance of centrality and prestige metrics, and a general conclusion can

Table 4: Authors ranked according to centrality measure

rank	Degree		Betweenness		Closeness	
1	Hsinchun Chen	59	Hsinchun Chen	89250.92	Hsinchun Chen	0.259
2	Edward A. Fox	55	Edward A. Fox	83163.92	Edward A. Fox	0.251
3	Terence R. Smith	31	Judith Klavans	57422.69	Judith Klavans	0.235
4	Carl Lagoze	31	William Y. Arms	52242.27	Gary Marchionini	0.234
5	Judith Klavans	27	Nina Wacholder	39226.5	Michael L. Nelson	0.229
6	Zan Huang	26	Craig Nevill-Manning	38808.08	Yiwen Zhang	0.226
7	Gary Marchionini	25	David M. Levy	35769.0	Ann M. Lally	0.226
8	William Y. Arms	21	Ann P. Bishop	32280.0	Lillian N. Cassel	0.226
9	Richard Furuta	21	Tobun D. Ng	30197.13	Byron Marshall	0.225
10	Luis Gravano	20	Gary Marchionini	29593.86	Rao Shen	0.225
11	Michael Freeston	19	Alexander Hauptmann	29142.0	William Y. Arms	0.224
12	Ian H. Witten	18	Catherine C. Marshall	28587.0	Anne Craig	0.221
13	Hector Garcia-Molina	18	Terence R. Smith	23691.87	Larry Brandt	0.221
14	Michael G. Christel	18	Carl Lagoze	22192.66	Terence R. Smith	0.219
15	David Millman	18	David Bainbridge	21168.03	Tobun D. Ng	0.219
16	Tamara Sumner	18	Michael L. Nelson	20696.41	James C. French	0.219
17	Diane Hillmann	18	Howard D. Wactlar	17577.0	Kurt Maly	0.212
18	Yilu Zhou	18	Ching-chih Chen	17309.67	Mohammad Zubair	0.212
19	Jialun Qin	18	John J. Leggett	15845.5	Hesham Anan	0.212
20	Mary Tiles	18	Elizabeth D. Liddy	14964.0	Xiaoming Liu	0.212

Table 5: Authors ranked according to PageRank/AuthorRank

Rank	PageRank	AuthorRank
1	Edward A. Fox	Hsinchun Chen
2	Hsinchun Chen	Edward A. Fox
3	Carl Lagoze	Ian H. Witten
4	Judith Klavans	Gary Marchionini
5	Richard Furuta	Hector Garcia-Molina
6	Gary Marchionini	Carl Lagoze
7	Michael G. Christel	Alexander G. Hauptmann
8	Terence R. Smith	Judith Klavans
9	Tamara Sumner	Richard Furuta
10	Ian H. Witten	Terence R. Smith
11	Alexander G. Hauptmann	Tamara Sumner
12	Hector Garcia-Molina	Ee-Peng Lim
13	Javed Mostafa	Michael G. Christel
14	Alexa T. McCray	Michael L. Nelson
15	Ee-Peng Lim	Wee Keong Ng
16	David Bainbridge	Javed Mostafa
17	Sally Jo Cunningham	David Bainbridge
18	Luis Gravano	J. Alfredo Sánchez
19	Catherine C. Marshall	Alexa T. McCray
20	W. Bruce Croft	Andreas Paepcke

be that no single measure is suited for all applications; each method has its virtues and utility [24, 9]. We verified and compared metrics in two ways: by the computation of the Spearman correlation coefficient across ranking methods, and by cross-validation against the dateset of JCDL program committee members.

The Spearman correlation coefficient is used to measure the strength of association between two variables. In our case, since betweenness generated only 153 authors with positive ranking, and closeness centrality has only been calculated for the largest component, we only compare degree centrality, PageRank, and AuthorRank. The correlation coefficient between the degree centrality and PageRank is 0.52, and the correlation coefficient between the degree centrality and AuthorRank is 0.30 (Figure 12). As expected, PageRank and AuthorRank are more closely correlated with a correlation coefficient of 0.75 (Figure 13).

We also verified each ranking method against a dataset consisting of all members of the JCDL, ADL and DL program committees from 1994 to 2004. This is meaningful, as program committee members are assumed to be prestigious actors in the co-authorship network. To that end, the names of all JCDL, ADL and DL program committee members were collected from the conference web sites or printed proceedings. Then, the highest scoring 50 authors for each ranking method (degree, closeness, betweenness, PageRank, AuthorRank) were matched one by one against each JCDL committee member to identify matches.

Figure 14 shows the result of this comparison. The highest ranking 5 authors for each method almost have a perfect match against the dataset of JCDL program committee members. Overall closeness ranking performs the worst, as only six authors of the 50 highest ranking authors are on the JCDL committees. This is not a surprise since closeness measures the distance to other authors, and since an author next to a prominent author is not necessarily also a prominent author. Degree centrality had mediocre performance.

Betweenness centrality performs the best among the three centrality measures. Since betweenness evaluates one's importance as a bridge between others, this suggests a committee member may be more likely to serve as a bridge between research groups than a non-committee member. The differences among betweeness, PageRank, and AuthorRank may not be statistically significant.

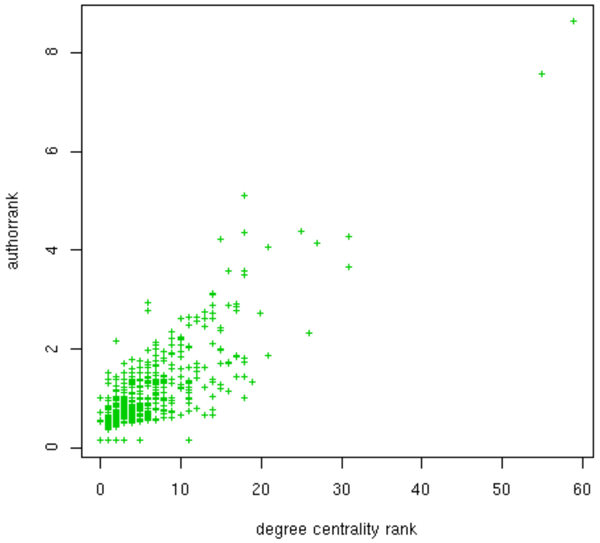

Figure 12: Scatter-plot of degree centrality vs. AuthorRank

5. CONCLUSIONS

In this paper we investigate the co-authorship network of the JCDL (and ADL/DL) conference series. We also present AuthorRank, an alternative metric for ranking authors' prestige in weighed co-authorship networks. So what does it all mean? What have we learned about the state of DL research 10 years after the first DL conference? Our data paints the picture of a domain that is in many ways still evolving the rich networks of collaboration common in other areas of the scientific enterprise. Our co-authorship graphs indicate a rich tapestry of collaborations across institutional boundaries, but demonstrate a significantly higher degree of clustering and dispersion than one would find in other domains. In comparison with other co-authorship networks for related disciplines, we find the JCDL co-authorship graph has a smaller largest component, a larger clustering coefficient and a larger characteristic path length. DL authors thus collaborate closely within specific clusters but restrict their collaborations to specific groups of interest.

This applies equally well to the matter of international collaboration. Approximately 72% of all authors are associated with US institutions, followed by a wide margin by Germany, the United Kingdom and Japan. The JCDL conference thus remains largely an American affair. More dire news is the amount of international collaboration. Only 7%

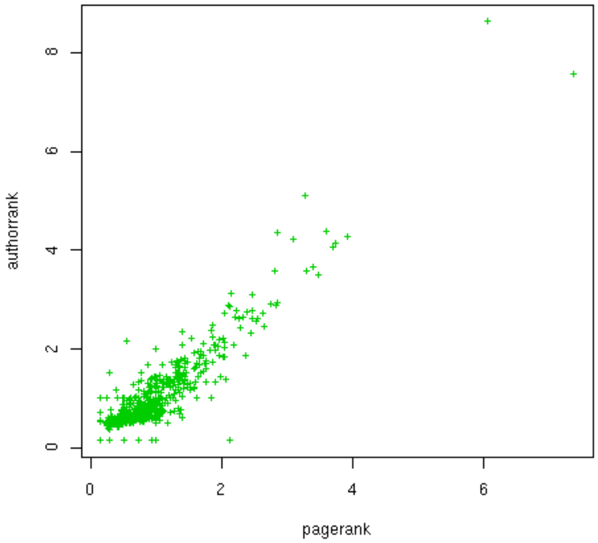

Figure 13: Scatter-plot of PageRank vs. AuthorRank

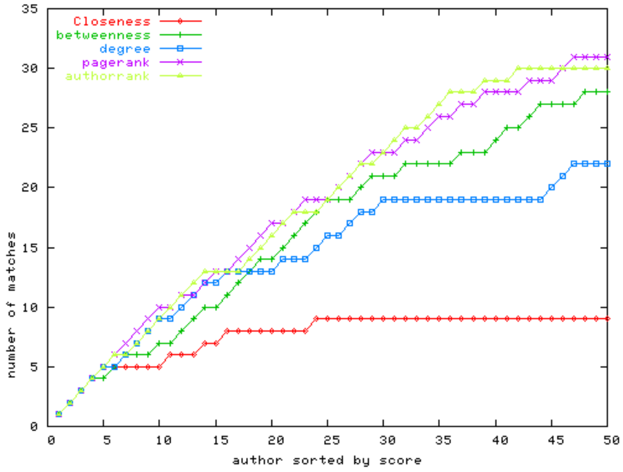

Figure 14: Ranking against JCDL program committee (1994-2004)

of all co-authorship relations concern international collaborations. This combined with the high degree of clustering and dispersion indicates the DL domain still leaves many opportunities for collaboration unexplored, both domestically and internationally.

Do these results mean collaboration is less valued in DL research? Of particular interest is our result demonstrating how well our calculations of author status, i.e. AuthorRank, in the co-authorship graph correspond to the JCDL program committees. Although the domain of DLs is less well-connected than other scientific domains, the value of collaboration still functions as an invisible hand guiding the selection of program committees in at least one seminal DL conference. It is thus of vital importance that a continued emphasis be placed on domestic and international collaboration to ensure DL research will be even more of the open, diverse, but well connected marketplace of ideas it is today.

6. ACKNOWLEDGEMENTS

The authors would like to thank Rick Luce, Linn Marks, Patrick Hochstenbach, and Jeremy Hussell from the LANL Research Library for their support of this work.

7. REFERENCES

[1] Country code domains. `http://www.iana.org/cctld/`.

[2] Graphviz - open source graph drawing software. `http://www.research.att.com/sw/tools/graphviz/`.

[3] The R project for statistical computing. `http://www.r-project.org/`.

[4] LANL CNLS 23th annual conference - networks: Structure, dynamics and function, 2003. `http://cnls.lanl.gov/Conferences/networks/`.

[5] A.-L. Barabási. *Linked-The new science of networks*. Perseus Publishing, Cambridge, MA, 2002.

[6] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998.

[7] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23:191–201, 2001.

[8] R. D. Castro and J. Grossman. Famous trails to Paul Erdos. *MATHINT: The Mathematical Intelligencer*, 21:51 – 63, 1999.

[9] S. Chakrabarti. *Mining the web*. Morgan Kaufmann Publishers, 2003.

[10] C. Chen and L. Carr. Trailblazing the literature of hypertext: author co-citation analysis (1989 - 1998). In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots*, pages 51–60. ACM Press, 1999.

[11] S. J. Cunningham and S. Dillon. Authorship patterns in information systems research. *Scientometrics*, 39(1):19 – 27, 1997.

[12] S. L. Esler and M. L. Nelson. Evolution of scientific and technical information distribution. *Journal of the American Society of Information Science*, 49:82–91, 1998.

[13] I. Farkas, I. Derenyi, H. Jeong, Z. Neda, Z. N. Oltvai, E. Ravasz, A. Schubert, A.-L. Barabasi, and T.Vicsek. Networks in life:scaling properties and eigenvalue spectra. *Physica A*, 314:25 – 34, 2002.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[15] M. Ley. DBLP computer science bibliography. `http://dblp.uni-trier.de/`.

[16] C. Li and G. Chen. Network connection strengths: Another power-law? Technical Report cond-mat/0311333, ArXiv, 2003.

[17] M. A. Nascimento, J. Sander, and J. Pound. Analysis of SIGMOD's coauthorship graph. *SIGMOD Record*, 32(3), 2003.

[18] M. Newman. Scientific collaboration networks. *Physical Review E.*, 64:016131, 2001.

[19] E. Otte and R. Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, pages 441–453, 2002.

[20] L. Page and S. Brin. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World-Wide Web Conference*, April 1998.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[22] A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sodring. Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36(2), 2002.

[23] B. Tjaden and et al. The Oracle of Bacon, 2003. `http://www.cs.virginia.edu/oracle/`.

[24] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[25] D. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, 2001.